BasB
Boletín de la Asociación Andaluza de Bibliotecarios

TABLAS DE CONTENIDO DE MONOGRAFIAS CON CARÁCTER COLECTIVO Y ENRIQUECIMIENTO DE REGISTROS BIBLIOGRAFICOS

Eduardo Peis
Universidad de Granada

RESUMEN

El concurso de las nuevas tecnologías ha incrementado enormemente las posibilidades de acceso a los recursos informativos. Paradójicamente, un porcentaje considerable de dichos recursos continúa siendo inaccesible al usuario. Parte de esta inaccesibilidad se evitaría enriqueciendo el contenido de los registros de la base de datos del catálogo en línea con información relativa a los trabajos intelectualmente individuales que conforman monografías colectivas. Este trabajo evalúa las tablas de contenido de este tipo de obras como fuente potencialmente eficaz para obtener términos significativos que representen los trabajos que contienen.

Palabras clave: Catálogos en línea / Enriquecimiento de registros bibliográficos / Monografías con carácter colectivo / Tablas de contenido / Términos significativos

INTRODUCCION

En algunas de las llamadas bibliotecas digitales se contempla la posibilidad de utilizar los denominados "libros dinámicos" (12). Se trata de documentos en los que se define una estructura básica que incluye todos los capítulos candidatos. El usuario tiene la capacidad de incluir los capítulos deseados y en el orden preferido. Los capítulos reales se obtienen como recursos digitales y de esta manera podría tratarse de la última versión, por lo que trasmitirían el conocimiento más actualizado. La tabla de contenidos y el índice de un "libro dinámico" podría incluir todos los capítulos candidatos, con indicaciones para distinguir los capítulos opcionales. Esta idea no es tan descabellada como podría parecer, algunos editores permiten ya la composición personalizada de algunos libros, en concreto manuales para la enseñanza, seleccionando los capítulos y su orden. Cada capítulo podría ser considerado como un recurso de información.

En el mundo de los libros impresos e incluso en el de la información electrónica lo anteriormente expuesto implicaría una cuestión intelectual: ¿es razonable tratar los elementos de un trabajo como piezas de información independientes?. En muchos casos no sería correcto este tratamiento debido a que un párrafo, e incluso un capítulo, puede no tener sentido si no es integrado en un trabajo como un todo. Es más, legalmente hablando (5), puede significar una afrenta a las intenciones del autor; la mayoría de los autores no piensan en sus libros como en una simple colección de parágrafos independientes.

No obstante, en el contexto de los libros impresos, se puede comprobar fácilmente que en toda colección bibliotecaria existen una serie de obras que a pesar de presentarse como un todo físico, contienen varios trabajos intelectualmente individuales. Es decir, cada capítulo o apartado de los que constituyen estas obras puede ser considerado aisladamente, al menos a efectos de recuperación de la información contenida en ellos, sin menoscabo de su inteligibilidad. La proporción que representan estas obras respecto al total de la colección no es en absoluto desdeñable. Basandose en su estudio sobre la colección de un colegio universitario, Hoffman y Magner (6) afirmaron que de alguna manera, aproximadamente un libro de cada cinco era un documento múltiple, una colección o una antología", y que "para cada libro de la estantería la biblioteca posee aproximadamente siete documentos". Desgraciadamente, pesar de las avanzadas posibilidades que presentan los catálogos en línea actuales, el acceso a estos trabajos que pueden ser considerados como intelectualmente individuales debe realizarse manualmente.

Por otra parte, la extraordinaria especialización y la velocidad de los cambios tecnológicos, hacen necesaria la rápida consulta de aspectos muy concretos y actualizados de un tema determinado. Mucha de la información de este tipo, además de estar vehiculada a través de publicaciones periódicas, está recogida en monografías con carácter colectivo. Las posibilidades para la recuperación y las necesidades informativas de los usuarios no han dejado de evolucionar y, por lo tanto, el papel del catálogo en línea continúa o debe continuar variando.

Para aquellos libros que realmente son colecciones de contribuciones separadas, añadir acceso a elementos de las tablas de contenido es equivalente a proporcionar acceso a nivel de artículo: proporciona acceso a trabajos intelectualmente individuales contenidos en un todo físico. Del mismo modo podríamos hablar de acceso a nivel de hoja para un conjunto de mapas y acceso a nivel de pieza a registros sonoros que contienen más de una unidad o a colecciones de partituras (3). De hecho, autores tan prestigiosos como Belkin y Saracevic (2) en un estudio para determinar principios de diseño de lo que denominaron la tercera generación de OPAC

(Open Public Access Catalog) destacaron la mejora de los registros documentales con la inclusión de la información de las tablas de contenido como deseable, no sólo con propósitos de recuperación sino también y quizás más importante, con propósitos de enjuiciar la relevancia.

Por lo tanto, una de las más importantes y acuciantes posibilidades de mejora del catálogo en línea podría ser la inclusión de información descriptiva que envíe a cada uno de los trabajos que constituyen una publicación cooperativa. Esta información descriptiva podría proceder de la tabla de contenidos.

En un importante trabajo sobre el uso del catálogo, formado por un conjunto de proyectos individuales coordinados por el *Centre for Catalogue Research* de la Universidad de Bath, una de las recomendaciones hacía hicapié en la necesidad de mejorar el acceso a los fondos más allá de la descripción bibliográfica tradicional. En sus propias palabras: "Las oportunidades que ofrecen los nuevos medios tecnológicos implican que debe incrementarse la explotación de los materiales bibliotecarios mediante el uso de palabras clave y frases de las páginas de título, las páginas de contenido y los índices..." (10).

Prabha y colaboradores (9), realizaron un análisis sobre el uso de libros de "no ficción", con el fin de estudiar el comportamiento de 331 usuarios de la principal biblioteca de la *Ohio State University* . Se contabilizaron las características o elementos de recuperación para la localización de los libros. Los resultados demostraron que, si exceptuamos el libro como un todo, la preferencia mostrada hacia la tabla de contenidos fue manifiesta (17,5%).

El incremento de información temática relativa a las monografías en los catálogos comenzó a prácticarse en la década de los setenta. El proyecto para el enriquecimiento temático de registros bibliográficos conocido como *SAP* (*Subject Access Project*), dirigido por Atherton (11), fue uno de los primeros y se puede considerar como punto de partida de otros como: el *SAP-Sweden* (14), el proyecto *EIS* (*Engineering Information System*) de la Universidad de Purdue (8), el *ESP* (*Enhancing Subject Project*) en la *ADFA* (*Autralian Defence Force Academy*) (1) o el proyecto *Mercury* de la Universidad Carnegie Mellon (7).

Todos estos proyectos, con mayores o menores variaciones metodológicas (empleo de fotocopias u originales, empleo de códigos para la selección o restricción del proceso a un tipo de material concreto) consistían en el establecimiento de una serie de sencillos criterios para la selección de términos significativos, sobre todo procedentes de los títulos de las tablas de contenido, y la inclusión de dichos términos en algún campo apropiado del registro MARC (13).

No obstante, estas experiencias no estuvieron exentas de limitaciones. En su momento, se planteó la cuestión de si el volumen de trabajos colectivos justificaría el desarrollo de métodos para acceder a sus elementos individuales. En este sentido, ya hemos comentado la investigación que llevaron a cabo Hoffman y Magner. Pues bien, de entre otros resultados encontraron que aproximadamente el 20% de los libros publicados anualmente eran documentos formados por múltiples trabajos y que, en la relativamente pequeña colección de 70.000 libros analizados, se contenían 517.300 trabajos, de los cuales 462.210 correspondían a libros cooperativos. Aproximadamente el 70% de los trabajos intelectualmente individuales estaban contenidos en unidades físicas colectivas. Estos trabajos podían considerarse invisibles al usuario ya que normalmente no eran accesibles a través del catálogo.

Además de los gastos adicionales y de las limitaciones metodológicas, aparecieron otra serie de cuestiones relacionadas con la sobrecarga y la pertinencia de los datos. ¿La inclusión de un tema o un nombre personal en la tabla de contenidos de un libro significa que el libro ofrece alguna información útil sobre ese tema o persona?. En algunas ocasiones, realmente no, cosa que puede ser fácilmente comprobada con la simple experiencia de consultar la tabla de contenidos de un libro y equiparar su contenido informativo con el contenido real del capítulo al que se refiere.

En su relativamente pequeño estudio sobre treinta y un libros de un tema específico, De Hart y Reitsma (4) encontraron que 383 de los 446 títulos de los capítulos (aproximadamente el 86%) "se juzgó que requerían un análisis en contexto para precisar su significado". Mientras que afirmaban que se obtenía un valor informativo significativo al añadir encabezamientos de capítulos a los registros bibliográficos, también reconocían que la inclusión de todos los encabezamientos podría, en muchos casos, generar un alto índice de ruido. No obstante, la abrumadora mayoría de libros que componían la muestra eran monografías física e intelectualmente unitarias. Incluso tratandose de monografías de tipo colectivo, si se empleasen los términos de las tablas de contenido para proporcionar acceso a los capítulos intelectualmente individuales que las componen, un número excesivo de términos no significativos, una gran cantidad de variantes gramaticales de una misma raiz léxica o un gran volumen de títulos con capacidad representativa nula, por poner algunos ejemplos, serían problemas que provocarían un incremento de complejidad operativa ya que incidirían sobre la efectividad de la recuperación.

El presente trabajo, mediante un análisis de la composición terminológica de las tablas de contenido, trata de demostrar que restringiendo dicho análisis a monografías "no-ficción", de tipo eminentemente científico y de carácter colectivo, aquel porcentaje de "no correspondencia" hallado por DeHart y Reitsma, disminuye enormemente y que los términos que componen estas tablas de contenido tienen por sí mismos una carga significativa, al menos suficiente para el acceso a los capítulos que representan.

Muestra y metodología

La población está compuesta por el subconjunto de monografías (sin contar con las series y

obras de referencia) añadidas durante el curso 1994-1995 a las colecciones de las bibliotecas de la Biblioteca Universitaria de Granada.

El total de 27.089 monografías incorporadas fue distribuido en cuatro estratos: humanidades, ciencias sociales, ciencias de la salud y ciencia y tecnología. Al seleccionar los trabajos incorporados a las bibliotecas representativas de cada estrato (15.545) y aplicar sobre estos un tamaño de muestra del 11 % obtuvimos una muestra de 1650 individuos.

Dado que de lo que se trataba era analizar características de las tablas de contenido como sujetos potencialmente tratables, debíamos obtener una nueva muestra conformada tan sólo por monografías de carácter colectivo. Para ello desarrollamos un proceso de selección sistemático aplicando un tamaño de muestra aproximado del 6 %. El resultado fue una segunda muestra consituida por cien elementos que compartían una serie de características comunes: son monografías de tipo colectivo; presentan la información sobre los trabajos individuales que contienen en formato tabla de contenidos; pertenecen a las áreas temáticas de documentación e informática aplicada; y, su incorporación (operativa) al sistema de catálogo en línea era reciente. Estaban representadas todas las categorías establecidas y en proporciones equivalentes a las halladas en una etapa analítica previa.

La aplicación de estratos obedeció al intento de seleccionar una muestra lo más representativa posible (ya que el volumen de trabajos incorporados a las diferentes colecciones varía enormemente) y a la necesidad de aplicar posteriormente algún método de inferencia estadística.

Se seleccionó un subconjunto de 50 elementos de esta muestra que se corresponden con monografías de temática relativa sobre todo a documentación-biblioteconomía e informática aplicada y, naturalmente, editadas en lenguas diferentes.

De cada uno de los elementos que conformaban este subconjunto se escogieron sistemáticamente los dos primeros y los dos últimos trabajos que contenían.

Los títulos de los doscientos trabajos resultantes fueron introducidos en una aplicación informática de gestión bibliográfica conformando la base de datos de prueba.

En un primer paso, se analizaron los títulos íntegros a fin de determinar su capacidad representativa, su tamaño y detectar si existen o no características que alteren dicha función representativa.

La aplicación informática a la que hemos aludido nos permitió posteriormente procesar los títulos a fin de aislar los términos que los componían. De igual manera, la transferencia de los datos a una hoja de cálculo informatizada posibilitó la tabulación de dichos datos y el cálculo de la frecuencia de aparición de dichos términos.

Se analizaron aspectos relativos a la cualidad representativa de los términos aislados (significativos o no) y a la existencia de categorías de términos que pueden ser considerados especiales respecto al tratamiento y la recuperación.

Resultados

Naturalmente, los resultados d ambos análisis están estrechamente relacionados.

Análisis de los títulos íntegros

Podemos constatar en la tabla siguiente que son pocos, tan sólo once de los doscientos títulos, los que aislados del todo al que pertenecen, no informan sobre el contenido real del trabajo al que designan. Es decir, el 5,5% de los títulos analizados fuera de la tabla de contenidos a la que pertenecen, no expresan claramente el tema o temas del trabajo al que intitulan.

TITULOS SIN CAPACIDAD REPRESENTATIVA FUERA DE CONTEXTO
Algunas ilustraciones sobre mi vida y obra
Epilogue
Generalità
The international scene
Introducción
Opening the horizon of expectations
Prólogo a la edición castellana

Relatoría general

Report on the afternoon panel discussion and open discussion

Results and further action

Roles and framework

La cifra anteriormente expresada puede parecer relativamente grande, sin embargo podemos refinar aún más este detalle analítico. Como se pondrá de manifiesto inmediatamente, al analizar término a término estos mismos títulos no representativos, se descubre que algunos de dichos términos pueden resultar adecuados y precisos para la recuperación de los trabajos de cuyos títulos forman parte.

Teniendo en cuenta lo anterior, el porcentaje de títulos no representativos, excesívamente generales o ambiguos se reduce hasta un 2,5% y, prácticamente el 100% de éstos se corresponden con capítulos con títulos del tipo de apéndice, *epilogue* , introducción o *results and further actions* , que ocupan en sus respectivas tablas de contenido una posición física inicial o final.

Aunque no se trata de una característica terminológica, es interesante determinar el tamaño de los títulos. El tamaño de los títulos abarca desde aquellos que están constituidos por una sola palabra (de 7 caracteres la más pequeña), hasta los que ocupan varias líneas impresas (el mayor con 5 líneas y un total de 29 términos; lo que, con caracteres de 10 puntos, un espacio en blanco entre palabras y con un espacio interlineal, representa un área aproximada de tres por doce centrímetros).

Son precisamente un mínimo de los títulos más amplios, no representativos y ambiguos los únicos en los que se ha desarrollado alguna figura del lenguaje con los términos que los componen. En efecto, fenómenos lingüísticos tales como sinonimias, anáforas o repeticiones, que podrían alterar la cualidad representativa a la que aludíamos anteriormente han sido detectados tan sólo en un 1,5% de los títulos aproximadamente.

Análisis de los términos

El siguiente paso es examinar los términos que conforman los títulos de la muestra. No es nuestra intención profundizar en el análisis hacia una evaluación detallada de la carga significativa de los términos, sino distinguir entre los términos (de cualquier tipo), que son útiles (a cualquier nivel) para representar el trabajo al que se refieren y aquellos que no son significativos (también denominados palabras vacías).

Una vez aislados, los términos de los doscientos títulos analizados anteriormente constituyen una nueva muestra para este segundo examen, que está formada por un total de 723 elementos. No obstante, esta cifra de 723 elementos es relativa a los términos sin contabilizar sus apariciones. Es decir, se han identificado 723 términos, pero en realidad, los doscientos títulos están formados por 1401 palabras.

El número medio de palabras por título sería por lo tanto de siete aproximadamente. Obviamente, alguno de los términos analizados se repiten diez, veinte y hasta más de cincuenta veces.

Efectivamente, si analizamos las frecuencias de aparición de los términos de los títulos, que aparecen en el siguiente gráfico, podemos observar cómo la inmensa mayoría, un total de 633 (lo que representa casi un 88%), aparecen tan sólo una vez y, en pocos casos, dos veces.

Existe un grupo intermedio de 76 términos, es decir el 10% del total, que presentan una frecuencia de aparición que se encuentra entre tres (la frecuencia más abundante de este grupo) y ocho apariciones (la más escasa).

Por contra, los 14 términos que representan el 2% del total y que formarían el tercer grupo, son los que se repiten con más frecuencia (desde 10 hasta 55 apariciones).

En otras palabras, el 88% de los términos (el primero de los grupos) representan el 52% del total de apariciones. Un 22,5% del conjunto de apariciones corresponde al grupo intermedio, constituido por 76 términos, mientras que el 25,5% de apariciones restante está formado por la repetición de tan sólo 14 de los términos analizados.

Se puede constatar cómo son precisamente algunas de las palabras vacias con respecto a la recuperación los que forman mayoritariamente este último grupo.

De hecho, si analizamos estos términos, comprobamos que constituyendo el 8,7% de los términos de la muestra, representan al 32,33% de las palabras que forman los títulos. Naturalmente se trata de artículos, preposiciones, conjunciones y algunos adverbios y pronombres. Trece de ellos, el 21%, forman el tercer grupo de términos de máxima aparición al que nos referíamos. En concreto, y por poner algunos ejemplos, la preposición "de" se repite 55 veces, los artículos "the" y "la" 43 y 34 veces respectivamente, la conjunción "and" aparece en 36 ocasiones, etcétera. No olvidemos, sin embargo, que dicho tercer grupo está formado por

catorce términos; el término catorceavo, " *information* ", que se repite 23 veces (a las que habría que sumar las ocho apariciones de variantes con la misma raiz léxica) no es una palabra no significativa; no obstante, al estar compuesta la muestra por trabajos de temática relativa a informatíca y documentación, se trata de un término del que se puede afirmar que posee, con relación a la recuperación, una capacidad discriminatoria muy baja.

Existe un grupo de términos (gráfico 2), que a pesar de su escasez pueden ser categorizados atendiendo a ciertas características semánticas que es conveniente que sean consideradas especiales, dado que potencialmente recibirán un tratamiento específico en el sistema. Se trata, en concreto, de nombres propios, nombres comerciales, fechas, acrónimos y topónimos.

Este conjunto supone el 9,5% de los términos de la muestra, pero representa escasamente el 6% de las apariciones. Por tipos, los términos de lugar son los más abundantes (32 palabras que aparecen en 37 ocasiones), mientras que los más escasos son los que hemos denominado nombres comerciales (5 términos con 8 apariciones).

Respecto a las apariciones por tipos (gráfico 3), podemos comprobar cómo los términos significativos (excluyendo las palabras vacías y algunos términos especiales), aunque constituyen sólo el 61% de las apariciones, corresponden al 81% del total de términos analizados.

Títulos analizados 200

Total términos 723

Total apariciones 1401

Más de la mitad de estos términos que hemos estimado como significativos se pueden agrupar formando lo que hemos llamado familias léxicas alrededor de una raiz común (tabla 6 en apéndices). Un total de 256 términos significativos, se agrupan en 90 familias léxicas, que suponen un total de 439 apariciones.

CONCLUSIONES

* Una proporción enorme (aproximadamente el 97,5 %) de los títulos de trabajos intelectualmente individuales que pertenen a monografías de carácter colectivo representan de forma adecuada su contenido. Son, por lo tanto, una fuente adecuada para la extracción de términos que posibiliten el enriquecimiento de la base de datos de un catálogo en línea, de forma que se procure acceso a dichos trabajos.

* Aquellos títulos que se pueden considerar como no representativos, excesívamente generales o ambiguos (tan sólo un 2,5 %), se corresponden mayoritariamente con capítulos cuyos títulos ocupan en sus respectivas tablas de contenido una posición física inicial o final, lo que podría facilitar su tratamiento.

* El tamaño físico de los títulos analizados es perfectamente procesable.

* Figuras del lenguaje tales como sinonimias, anáforas o repeticiones, que podrían alterar la cualidad representativa de los títulos analizados han sido detectados en un porcentaje ínfimo. No obstante, el "ruido" que podrían provocar tales figuras es perfectamente soportable y puede ser atenuado mediante el tratamiento al que se somete a los términos individuales.

* Se obtienen aproximadamente siete palabras por título procesado por lo que, con relación a los registros enriquecidos mediante este proceso, se puede decir que idealmente se incrementa el número de elementos temáticos que contienen estos registros en un 700%. Decimos idealmente por que muchos de estos términos se repiten, son palabras vacías o se trata de derivaciones gramaticales de una misma raiz léxica.

* Es escaso el número de palabras vacías con relación a la recuperación. No obstante, éstas son también los términos que, en proporción, suponen un mayor número de apariciones, lo que obviamente implica que son también los que más se repiten. La configuración y aplicación de un eficaz antidiccionario evita esta dificultad y, aunque este proceso reduce el número de términos útiles para la recuperación procedentes de los títulos, incrementa la cualidad representativa de los que restan.

* Un pequeño porcentaje de términos pueden ser considerados especiales, ya que se trata de nombres propios, fechas, acrónimos, topónimos y nombres comerciales. No obstante, no son especiales en cuanto a su tratamiento, ya que resultan significativos y muy precisos para la recuperación.

* La adecuación de los títulos de este tipo de trabajos para la extracción de palabras clave para la recuperación, puede ser refrendada por el alto porcentaje de términos procedentes de dichos títulos que son significativos; y, aunque más de la mitad de estos términos se agrupan formando familias léxicas alrededor de una raiz común, el alto porcentaje de apariciones que presentan, con una tasa de repetición baja, y la posibilidad de aplicar un algoritmo para el *stemming* , los convierte en especialmente útiles.

REFERENCIAS BIBLIOGRAFICAS

1. BEATTY S. ESP at ADFA after five years. En: Cataloging for On-line. Access *National Cataloguing Conference,* (*Melbourne, Ormond College University, 28-30 Nov. 1991*), págs. 126-140. Melbourne: University, 1991.
2. BELKIN, N. J. y SARACEVIC, T. "Desing Principles for Third-Generation Online Public Access Catalogs: Taking Account of Users and Library Use", *Annual Review of OCLC Research* , (1992), vol. July 1991-June 1992, págs. 43-45.
3. BOSS, R. W. "Online Catalog functionality in the 90's: vendor responses to a model RFP", *Library Technology Reports* , (1993), vol. 29 - nº 5, págs. 587-618.
4. DE HART, F. E. y REITSMA, R. "Subject Searching and Tables of Contents in Single-Work Titles", *Technical Services Quaterly* , (1989), vol. 7, págs. 39-47.
5. FERNANDEZ-MOLINA, J. C. y PEIS, E. Los derechos morales del autor en un entorno electrónico. En: *VI Jornadas Españolas de Documentación Automatizada (Valencia, 28-31 de octubre de 1998)* . Valencia, Fesabid, 1998.
6. HOFFMAN, H. H. y MAGNER, J. L. "Future outlook: better retrieval trough analitic catalogs", *The Journal of Academic Librarianship* , (1985), vol. 11 - nº 2, págs. 152-167.
7. MICHALAK, T. J. "An Experiment in Enhancing Catalog Records at Carnegie Mellon University", *Library Hi Tech* , (1990), vol. 3 - nº 31, págs. 33-41.
8. POSEY, E. D. y ERDMANN, C. A. "An Online UNIX-Based Engineering Library Catalog: Purdue University Engineering Library", *Science & Technology Libraries* , (1986), vol. 32 - nº 6:4, págs. 31-43.
9. PRABHA, C. G., RICE, D. y CAMERON, D. *Nonfiction Book Use by Academic Library Users* , Dublin, OH, OCLC, 1988.
10. SEAL, A., P. BRYANT y C. HALL *Full and short entry catalogues: library needs and uses* , Aldershot, Gower, 1982.
11. SETTEL B, ATHERTON P. Augmenting Subject Description for Books in Online Catalogs. En: Atherton P, ed. Redesing of Catalogs and Indexes for Improved Online Subject Access: Selected Papers of Pauline Atherton. Phoenix: Oryx Press, 1985.
12. WIEDERHOLD, G. "Digital Libraries, value, and productivity", *Communications of the ACM* , (1995), vol. 38 - nº 4, págs. 85-96.
13. WITTENBACH SA. Building a Better Mousetrap: Enhanced Cataloging and Access for the Online Catalog. En: Ra M, ed. Advances in Online Public Access Catalogs. Vol. 1. Westport, CT: Meckler, 1992:74-92.
14. WORMELL, I. *Subject access project: SAP: Improved subject retrieval for monographic publications* , Lund, Lund University, 1985.

NOTA: los gráficos citados deben ser consultados en la versión impresa.

C/ Olleríias, 45-47, 3° D | 29012 - MÁLAGA | Tel 952 213 188 |Fax 952 604 529 | Correo-e: aab@aab.es